American Journal of Epidemiology
© The Author 2010. Published by Oxford University Press on behalf of the Johns Hopkins Bloomberg School of Public Health. This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.5), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
Vol. 172, No. 9
DOI: 10.1093/aje/kwq224
Advance Access publication: August 28, 2010

Practice of Epidemiology

Statistical Criteria for Selecting the Optimal Number of Untreated Subjects Matched to Each Treated Subject When Using Many-to-One Matching on the Propensity Score

Peter C. Austin*

* Correspondence to Dr. Peter C. Austin, Institute for Clinical Evaluative Sciences, G1 06, 2075 Bayview Avenue, Toronto, Ontario, Canada M4N 3M5 (e-mail: peter.austin@ices.on.ca).

Initially submitted April 21, 2010; accepted for publication June 18, 2010.

Propensity-score matching is increasingly being used to estimate the effects of treatments using observational data. In many-to-one (*M*:1) matching on the propensity score, *M* untreated subjects are matched to each treated subject using the propensity score. The authors used Monte Carlo simulations to examine the effect of the choice of *M* on the statistical performance of matched estimators. They considered matching 1–5 untreated subjects to each treated subject using both nearest-neighbor matching and caliper matching in 96 different scenarios. Increasing the number of untreated subjects matched to each treated subject tended to increase the bias in the estimated treatment effect; conversely, increasing the number of untreated subjects matched to each treated subject decreased the sampling variability of the estimated treatment effect. Using nearest-neighbor matching, the mean squared error of the estimated treatment effect was minimized in 67.7% of the scenarios when 1:1 matching was used. Using nearest-neighbor matching or caliper matching, the mean squared error was minimized in approximately 84% of the scenarios when, at most, 2 untreated subjects were matched to each treated subject. The authors recommend that, in most settings, researchers match either 1 or 2 untreated subjects to each treated subject when using propensity-score matching.

bias (epidemiology); matching; Monte Carlo method; observational study; propensity score

Abbreviations: ANOVA, analysis of variance; ATT, average treatment effect for the treated; MSE, mean squared error.

In an observational study of the effects of treatments, exposures, or interventions, the propensity score is the probability of treatment assignment conditional on observed baseline covariates (1–4). In the absence of unmeasured confounding, conditioning on the propensity score allows one to obtain unbiased estimates of average treatment effects (1). Propensity-score matching is used frequently in the medical literature (5–7). The most common implementation of propensity-score matching is 1:1 matching, in which pairs of treated and untreated subjects are formed. The effect of treatment may be estimated by directly comparing outcomes between treated and untreated subjects in the matched sample.

Anecdotally, a criticism of propensity-score matching by clinical investigators and medical journal editors and reviewers is that it is ''wasteful'' of sample size. For instance, imagine a sample consisting of 1,000 subjects, of whom 100 were treated. The effect of treatment would be estimated in the matched sample consisting of, at most, 200 subjects. The remaining 800 unmatched, untreated subjects are apparently ''wasted,'' since they are not used in estimating the effect of treatment on outcomes. Several studies in the medical literature have used many-to-one (*M*:1) matching on the propensity score (8–18). Using this approach, M ($M > 1$) untreated subjects are matched to each treated subject.

Our objective in the current paper is to develop criteria for selecting the number of untreated subjects to match to each treated subject to optimize estimation of treatment effects when *M*:1 matching is used. First, we present a conceptual framework for propensity-score matching. We describe why discarding unmatched, untreated subjects does not introduce bias into the estimate of treatment effect. We then use Monte

Downloaded from https://academic.oup.com/aje/article/172/9/1092/147493 by guest on 18 April 2024

Carlo simulations to determine the number of untreated subjects to match to each treated subject in order to optimize estimation of treatment effects. Finally, we summarize our findings and discuss them in the context of the literature.

A CONCEPTUAL FRAMEWORK FOR PROPENSITY-SCORE MATCHING

In the potential outcomes framework proposed by Rubin (19), each subject has a pair of potential outcomes: $Y_i(0)$ and $Y_i(1)$, the outcomes under the control and active treatments, respectively. However, each subject receives only the control or the active treatment. Let Z be an indicator variable denoting the actual treatment received ($Z = 0$ for control treatment vs. $Z = 1$ for active treatment). Thus, only 1 outcome, Y, is observed for each subject: the outcome under the actual treatment received. Y_i is defined to be equal to $Y_i(0)$ if $Z_i = 0$ and to be equal to $Y_i(1)$ if $Z_i = 1$ (alternatively, $Y = ZY(1) + (1 - Z)Y(0)$). The average treatment effect for the treated (ATT), defined as $E[Y(1) - Y(0)|Z = 1]$, is the average effect of treatment on those subjects who ultimately received the treatment (20). Propensity-score matching allows one to estimate the ATT (20).

Comparing outcomes between treated and untreated subjects in a sample in which treatment assignment is not confounded with either measured or unmeasured baseline covariates allows one to obtain an unbiased estimate of the treatment effect. When there are no systematic differences in measured or unmeasured baseline covariates between a sample of treated subjects and a sample of untreated subjects, any difference in outcomes can be attributed to the effect of treatment. When estimating the ATT using observational data, one requires a sample of untreated subjects such that there are no observed systematic differences between the sample of treated subjects and the sample of untreated subjects. Then, in the absence of unmeasured confounding, an unbiased estimate of the ATT can be obtained by comparing outcomes between treated and untreated subjects in the matched sample (1). The initial sample of untreated subjects serves only as a pool of potential controls from which to find appropriate matches for treated subjects. When using propensity-score matching to form a matched sample in which observed systematic differences between treated and untreated subjects have been minimized, it is irrelevant how many untreated subjects were discarded. The important issue is that one has created a sample in which treatment selection is not confounded with measured baseline covariates. Then, under the assumption of no unmeasured confounding, one can estimate the ATT. From a conceptual perspective, the fact that only a minority of untreated subjects may have been used does not affect the estimate of the ATT.

While M:1 matching on the propensity score is not necessary from a conceptual perspective, there may be practical reasons for adopting this strategy. Increasing the number of untreated subjects included in each matched set may increase the precision of the estimated treatment effect. The optimum number of untreated subjects needed to match to each treated subject probably reflects the traditional variance-bias trade-off: Increasing the number of untreated subjects matched to each treated subject will increase the size of the matched sample, probably resulting in estimates of treatment effect with increased precision. However, increasing the number of untreated subjects matched to each treated subject may result in the matching of increasingly dissimilar subjects. This may increase bias in estimating the effect of treatment. Our objective in this paper is to determine the optimal number of untreated subjects to match to each treated subject when M:1 matching on the propensity score is used.

MONTE CARLO SIMULATIONS

We conducted an extensive series of Monte Carlo simulations to examine the impact of increasing the number of untreated subjects matched to each treated subject on the estimation of treatment effects.

Methods

Our Monte Carlo simulations used a design similar to those that have been described elsewhere (21–26). We used a complete factorial design in which the following factors were allowed to vary: the sample size of each simulated data set, the proportion of subjects within each simulated data set who were treated, the strength of the relation between baseline covariates and the log odds of the probability of receiving the treatment, and the strength of the relation between baseline covariates and the outcome.

We randomly generated data sets of 4 different sizes: 500, 1,000, 5,000, and 10,000 subjects per simulated data set. For each subject, we randomly generated 5 baseline covariates (x_1–x_5) from independent standard normal distributions. We assumed that the probability of treatment assignment (P_{treat}) was related to these baseline covariates via the following logistic regression model:

$$\log\left(\frac{P_{\text{treat}}}{1 - P_{\text{treat}}}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5.$$

A dichotomous variable (Z) denoting treatment status was then generated for each subject from a Bernoulli distribution with subject-specific parameter P_{treat}. The value of β_0 was fixed so that approximately the desired proportion of the subjects would receive the treatment, with the remaining subjects being untreated. We considered 4 different proportions of treated subjects: 0.02, 0.05, 0.10, and 0.15. We considered 2 different sets of regression coefficients relating the baseline covariates to the log odds of treatment: In the weak covariate scenario, the regression coefficients β_1, β_2, β_3, β_4, and β_5 took the values log(1.5), log(2), log(3), log(4), and log(5), respectively. In the strong covariate scenario, the regression coefficients took on the values log(2), log(2), log(5), log(5), and log(10) respectively.

A continuous outcome was generated for each subject using the following linear model:

Downloaded from https://academic.oup.com/aje/article/172/9/1092/147493 by guest on 18 April 2024

$$Y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \alpha_4 x_4 + \alpha_5 x_5 + Z + \varepsilon,$$

where $\varepsilon \sim N(0,\sigma^2)$. We fixed the values of $\alpha_0, \ldots, \alpha_5$ to be equal to 0, 1.5, 2, 3, 4, and 5, respectively. The value of σ was set so that, among untreated subjects, variation in the baseline covariates explained 2%, 13%, or 26% of the variation in the response variable (these values of R^2 have been described by Cohen (27) as weak, moderate, and strong effect sizes). Thus, we considered 96 different scenarios: 4 sample sizes $\times$ 4 proportions of subjects in the data set who were treated $\times$ 2 sets of regression coefficients relating baseline covariates to treatment $\times$ 3 sets of R^2 for the magnitude of the effect of baseline covariates on the outcome.

Once a data set had been randomly generated, we constructed a series of matched samples using 2 different matching methods: nearest-neighbor matching using the logit of the propensity score and nearest-neighbor matching within specified calipers using the logit of the propensity score. In each case, the propensity score was estimated using a logistic regression model to regress the treatment indicator variable (Z) on the baseline covariates (x_1–x_5). The estimated propensity score was the predicted probability of treatment assignment derived from the fitted logistic regression model.

The first matching method was nearest-neighbor matching without replacement using the logit of the propensity score (28, 29). We first used 1:1 matching in which a single untreated subject was matched to each treated subject. The treated subjects were randomly ordered. The first treated subject was selected and was matched to the untreated subject whose logit of the propensity score was closest to that of the treated subject. Since we were matching without replacement, matched untreated subjects were not available to serve as matches for subsequent treated subjects. This process was repeated until a match was found for each treated subject. We then had a matched sample consisting of pairs of treated and untreated subjects. We repeated the above process to create matched samples using 2:1, 3:1, 4:1, and 5:1 matching on the propensity score. Note that because we are using nearest-neighbor matching, a sufficient number of unmatched subjects will be available for each treated subject, since this method does not require that the difference in the logit of the propensity score between treated and untreated subjects within the same matched set be within a prespecified maximum distance.

The second matching method was nearest-neighbor matching using specified calipers without replacement (3). This approach is similar to that described above, with 1 exception: The difference in the logit of the propensity score between treated and untreated subjects in the propensity-score-matched set was required to be less than a prespecified maximum. We used a caliper of width equal to 0.2 of the standard deviation of the logit of the propensity score. This caliper width was selected because it (or one close to it) was shown in recent research to result in optimal estimation compared with other choices of caliper (30). Because of the imposition of the constraint that the logit of the propensity score of matched subjects could differ by, at most, a fixed amount, it is possible that insufficient numbers of untreated subjects will be available for matching to some treated subjects. Thus, when using M:1 matching ($M > 1$), it is conceivable that, while some matched sets will contain M untreated subjects, some matched sets will contain fewer than M untreated subjects.

Once a matched sample had been constructed, we estimated the treatment effect by the difference in the mean outcome between treated and untreated subjects in the matched sample. When estimating the mean outcome in the matched untreated subjects, each untreated subject was weighted by the reciprocal of the number of untreated subjects within the matched set to which that untreated subject belonged (20). The balance in the 5 baseline covariates between treated and untreated subjects was assessed using the absolute standardized difference (31), which was estimated using a method that accounted for the M:1 matched nature of the samples (32).

The above simulations and analyses were repeated 1,000 times. The mean estimated treatment effect and the mean absolute standardized differences were computed across the 1,000 simulated data sets. We also estimated the variance of the estimated treatment effect across the 1,000 simulated data sets. Finally, the mean squared error (MSE) of the estimated treatment effect was computed across the 1,000 simulated data sets. The MSE is equal to the sum of the variance of an estimator and the square of the bias of the estimator (33).

Results

Bias. Web Figure 1 (which is posted on the *Journal*'s Web site (http://aje.oxfordjournals.org/)) illustrates the relation between M (the number of untreated subjects matched to each treated subject) and bias in estimating the true treatment effect. For each scenario, we have used a ''filled-in'' plotting symbol to identify the value of M that minimized bias (for improved clarity, results for $N = 5{,}000$ are not displayed in any of the Web figures). When caliper matching was used, bias was minimized in 87 (90.6%) of the 96 scenarios when 1:1 matching was employed. In general, bias increased as M increased with M:1 matching. When nearest-neighbor matching was used, bias was minimized in all 96 scenarios when 1:1 matching was employed.

Sampling variance of the estimated treatment effect. Web Figure 2 shows the relation between M and the sampling variance of the estimated treatment effect. When caliper matching was used, the sampling variance of the estimated treatment effect was minimized in 46 of the 96 scenarios when 2:1 matching was employed; in 46 of the 96 scenarios when 3:1 matching was employed; in 3 of 96 scenarios when 4:1 matching was employed; and in 1 of the 96 scenarios when 5:1 matching was employed. When nearest-neighbor matching was used, the sampling variance was minimized in 7 of the 96 scenarios when 3:1 matching was employed; in 14 of the 96 scenarios when 4:1 matching was employed; and in 75 of the 96 scenarios when 5:1 matching was employed. When nearest-neighbor matching was used, sampling variances tended to decrease with increasing M.

MSE of the estimated treatment effect. Web Figures 3–8 illustrate the relation between M and the MSE of the

Downloaded from https://academic.oup.com/aje/article/172/9/1092/147493 by guest on 18 April 2024

estimated treatment effect. As with the previous plots, a solid-color plotting symbol was used to identify the value of M that minimized MSE within a given scenario.

For each of the 96 scenarios, the value of M that minimized MSE was determined. When caliper matching was used, the median value of M that minimized MSE was 2 across the 96 scenarios (the 25th and 75th percentiles were 1 and 2, respectively). MSE was minimized in 26 (27.1%), 54 (56.3%), 15 (15.6%), 1 (1.0%), and 0 (0%) scenarios when M was equal to 1, 2, 3, 4, and 5, respectively. When nearest-neighbor matching was used, the median value of M that minimized MSE was 1 across the 96 scenarios (25th percentile = 1; 75th percentile = 2). MSE was minimized in 65 (67.7%), 16 (16.7%), 7 (7.3%), 3 (3.1%), and 5 (5.2%) scenarios when M was equal to 1, 2, 3, 4, and 5, respectively. The values of N, P_{treat}, strong/weak association, and R^2 for scenarios in which 5:1 matching resulted in estimates with the lowest MSE were 1) 1,000, 2, weak, 0.02; 2) 1,000, 2, strong, 0.02; 3) 500, 2, weak, 0.02; 4) 500, 2, strong, 0.02; and 5) 500, 5, weak, 0.02, respectively.

An analysis of variance (ANOVA) model was used to examine the association between the 4 factors of the Monte Carlo study and the value of M that minimized MSE. This was done separately for caliper matching and nearest-neighbor matching. The initial ANOVA model included all 4 main effects and all 2-way interactions between main effects. We then excluded all 2-way interactions that were not significant ($P > 0.05$) in the original ANOVA model; we report the results based on the model with all 4 main effects and those interactions that were statistically significant.

When using caliper matching, the main effects for sample size ($P < 0.0001$), P_{treat} ($P < 0.0001$), and R^2 ($P < 0.0001$), along with the 2-way interaction between P_{treat} and sample size ($P = 0.0214$), were significantly associated with the value of M that minimized MSE. The value of M that minimized MSE decreased as P_{treat} increased. In general, when P_{treat} and R^2 were fixed, the value of M that minimized MSE tended to decrease as the sample size increased.

When using nearest-neighbor matching, the main effects for sample size ($P < 0.0001$), P_{treat} ($P < 0.0001$), a strong versus weak association between covariates and treatment selection ($P = 0.0142$), and R^2 ($P < 0.0001$), along with the 2-way interactions between N and P_{treat} ($P < 0.0001$), N and R^2 ($P < 0.0001$), and P_{treat} and R^2 ($P < 0.0001$), were significantly associated with the value of M that minimized MSE. The value of M that minimized MSE was, on average, lower when the baseline covariates were strongly associated with treatment selection. When P_{treat} and R^2 were fixed, in many settings the value of M that minimized MSE decreased as N increased. In some settings, a U-shaped relation between N and the value of M that minimized MSE was observed. In a minority of settings, other relations were observed. When N and R^2 were fixed, the value of M that minimized MSE either decreased as P_{treat} increased or had a U-shaped relation with P_{treat}. Finally, when N and P_{treat} were fixed, in 75% of the scenarios the value of M that minimized MSE decreased with increasing values of R^2. In the remaining scenarios, either the relation was U-shaped or M increased with increasing R^2.

The effect of the number of untreated subjects matched to each treated subject on covariate balance in the matched sample is shown in Web Figures 9 and 10 for nearest-neighbor matching and caliper matching, respectively. (We report results for only the settings with $R^2 = 0.02$, since the results were identical regardless of the value of R^2. This is due to the standardized difference being independent of the outcome—it depends only on baseline covariates.) With nearest-neighbor matching, we observed that imbalance tended to increase as M increased. In many settings, the relation between the absolute standardized difference and M was approximately linear. For some covariates, mean standardized differences exceeded 100% in some settings when 5:1 matching was used. Covariate imbalance in the matched sample tended to be lower when caliper matching was used in comparison with nearest-neighbor matching. When using caliper matching, the mean absolute standardized difference was less than 32% in all scenarios examined, regardless of the choice of M. With caliper matching, there were some scenarios in which modest improvements in balance were observed with 2:1 matching in comparison with 1:1 matching.

DISCUSSION

In our Monte Carlo simulations, we found that, on average, increasing the number of untreated subjects matched to each treated subject increased the bias of the estimated treatment effect; conversely, it tended to result in increased precision. When using nearest-neighbor matching, we found that MSE was minimized in 67.7% of the 96 scenarios when 1 untreated subject was matched to each treated subject. For either matching method, MSE was minimized in at least 84% of the scenarios when either 1 or 2 untreated subjects were matched to each treated subject. These findings suggest that in the majority of settings, using 1:1 or 2:1 matching will result in optimal estimation of treatment effects when employing fixed M:1 matching. When using caliper matching, we observed that the optimal value of M tended to decrease as the proportion of treated subjects increased.

There is a paucity of explicit research into the effect of increasing the number of untreated subjects matched to each treated subject. Imbens suggested that ''within the class of matching estimators, using only a single match leads to the most credible inference with the least bias, at most sacrificing some precision'' (20, p. 14). The results of our extensive Monte Carlo simulations provided confirmation of this suggestion: We observed increased bias as the number of untreated subjects matched to each treated subject increased. Furthermore, when using nearest-neighbor matching, MSE was minimized in 67.7% of the scenarios when 1:1 matching was employed.

Rosenbaum and Rubin (28) examined the bias due to incomplete and inexact matching when matching treated and untreated subjects on a set of baseline covariates. Incomplete matching occurs when there are treated subjects for whom no appropriate untreated subjects are identified. Inexact matching occurs when a treated subject and an untreated subject whose covariates are not identical are matched. In

Downloaded from https://academic.oup.com/aje/article/172/9/1092/147493 by guest on 18 April 2024

an empirical example, Rosenbaum and Rubin show that the bias due to incomplete matching can be substantial (28). Thus, in conventional matching, an important issue is not whether there are unmatched untreated subjects; rather, a much more important issue relates to whether there are unmatched treated subjects. Rosenbaum and Rubin suggest that, rather than use exact matching and risk bias due to incomplete matching, one can match using a multivariate nearest-neighbor method (such as the propensity score) and thus avoid biases due to incomplete bias (28). The resultant cost is only minor bias due to inexact matching. In the current study, estimation using nearest-neighbor matching would not have suffered from incomplete matching bias because sufficient matches were found for all treated subjects, since no constraints were placed upon the maximum difference in propensity score between treated and untreated subjects in the same matched set. However, estimation using caliper matching may have suffered from incomplete matching bias to a limited extent, since no matches may have been found for some treated subjects due to the constraint that the difference in the logit of the propensity score between treated and untreated subjects was required to not exceed a maximal value.

In the current study, we examined the impact of the number of untreated subjects matched to each treated subject in the context of propensity-score matching. The issue of how many subjects to include in a matched set has received greater attention in case-control studies. In case-control studies, cases (subjects who experience the outcome of interest) are matched with controls (subjects who did not experience the outcome of interest). Ury demonstrated that ''the theoretical efficiency of a 1:M case-control ratio for estimating a relative risk of about 1, relative to having complete information on the control population ($M = \infty$), is $M/(M + 1)$. Thus, 1 control per case is 50% efficient, while 4 per case is 80% efficient'' (35, p. 169). Thus, in case-control studies, increasing the number of controls matched to each case results in improved efficiency; however, the relative gains in efficiency are minor once M exceeds 5 or so. In contrast, when using propensity-score matching, there is a trade-off between bias and variance that does not exist in case-control studies. We have shown that, in many settings, the trade-off can be optimized by matching either 1 or 2 untreated subjects to each treated subject. In only a very small minority of settings was using 5 untreated subjects per case optimal.

We have demonstrated that increasing the number of untreated subjects matched to each treated subject can result in increased bias in estimating treatment effects. However, there are additional limitations to having more than 1 untreated subject matched to each treated subject. In particular, it can make estimation of the variance of the estimated treatment effect more difficult. For instance, when outcomes are binary, McNemar's test can be used to compare the proportion of successes between the 2 treatment groups when 1:1 matching is employed. However, when multiple untreated subjects are matched to each treated subject, it is unclear how the statistical significance of the risk difference should be determined.

We have examined criteria for determining the optimal number of untreated subjects to match to each treated subject when using fixed M:1 matching on the propensity score. There are alternatives to M:1 matching that we have not examined in the current paper because of space constraints. Ming and Rosenbaum (36) demonstrated that matching with a variable number of controls can reduce bias substantially in comparison with matching with a fixed number of controls. Furthermore, we have not considered full matching, in which multiple untreated subjects are matched to each treated subject or multiple treated subjects are matched to each treated subject, resulting in all subjects being included in a matched set (37–39). Full matching may often have superior performance to fixed M:1 matching (38). In this study, we have focused on M:1 matching because of its more frequent use in the medical literature.

In summary, we recommend that, in most settings, researchers match either 1 or 2 untreated subjects to each treated subject when using fixed M:1 matching on the propensity score. Using only 1 untreated subject for each treated subject will tend to minimize bias. In some settings, attempting to match 2 untreated subjects to each treated subject will result in improved precision without a commensurate increase in bias.

ACKNOWLEDGMENTS

Author affiliations: Institute for Clinical Evaluative Sciences, Toronto, Ontario, Canada; and Department of Health Management, Policy and Evaluation, Faculty of Medicine, University of Toronto, Toronto, Ontario, Canada.

This study was supported by the Institute for Clinical Evaluative Sciences, which is funded by an annual grant from the Ontario Ministry of Health and Long-Term Care. This study was also supported in part by an operating grant from the Canadian Institutes of Health Research (funding no. MOP 86508). Dr. Austin was supported in part by a Career Investigator Award from the Heart and Stroke Foundation of Ontario.

The opinions, results, and conclusions reported in this paper are those of the author and are independent from the funding sources. No endorsement by the Institute for Clinical Evaluative Sciences or the Ontario Ministry of Health and Long-Term Care is intended or should be inferred.

Conflict of interest: none declared.

REFERENCES

1. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70(1):41–55.
2. Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. *J Am Stat Assoc.* 1984;79(387):516–524.
3. Rosenbaum PR, Rubin DB. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *Am Stat*. 1985;39(1):33–38.
4. Austin PC, Mamdani MM. A comparison of propensity score methods: a case-study estimating the effectiveness of post-AMI statin use. *Stat Med*. 2006;25(12):2084–2106.

Downloaded from https://academic.oup.com/aje/article/172/9/1092/147493 by guest on 18 April 2024

5. Austin PC. A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Stat Med*. 2008;27(12):2037–2049.
6. Austin PC. Propensity-score matching in the cardiovascular surgery literature from 2004 to 2006: a systematic review and suggestions for improvement. *J Thorac Cardiovasc Surg*. 2007;134(5):1128–1135.
7. Austin PC. Primer on statistical interpretation or methods report card on propensity-score matching in the cardiology literature from 2004 to 2006: a systematic review. *Circ Cardiovasc Qual Outcomes*. 2008;1(1):62–67.
8. Boening A, Friedrich C, Hedderich J, et al. Early and medium-term results after on-pump and off-pump coronary artery surgery: a propensity score analysis. *Ann Thorac Surg*. 2003; 76(6):2000–2006.
9. Aronow HD, Novaro GM, Lauer MS, et al. In-hospital initiation of lipid-lowering therapy after coronary intervention as a predictor of long-term utilization: a propensity analysis. *Arch Intern Med*. 2003;163(21):2576–2582.
10. Magee MJ, Jablonski KA, Stamou SC, et al. Elimination of cardiopulmonary bypass improves early survival for multivessel coronary artery bypass patients. *Ann Thorac Surg*. 2002;73(4):1196–1202.
11. Sernyak MJ, Desai R, Stolar M, et al. Impact of clozapine on completed suicide. *Am J Psychiatry*. 2001;158(6):931–937.
12. Chukwuemeka A, Weisel A, Maganti M, et al. Renal dysfunction in high-risk patients after on-pump and off-pump coronary artery bypass surgery: a propensity score analysis. *Ann Thorac Surg*. 2005;80(6):2148–2153.
13. Reeves BC, Ascione R, Caputo M, et al. Morbidity and mortality following acute conversion from off-pump to on-pump coronary surgery. *Eur J Cardiothorac Surg*. 2006;29(6): 941–947.
14. Rajakaruna C, Rogers CA, Angelini GD, et al. Risk factors for and economic implications of prolonged ventilation after cardiac surgery. *J Thorac Cardiovasc Surg*. 2005;130(5): 1270–1277.
15. Kaw R, Golish J, Ghamande S, et al. Incremental risk of obstructive sleep apnea on cardiac surgical outcomes. *J Cardiovasc Surg (Torino)*. 2006;47(6):683–689.
16. Ahmed A, Perry GJ, Fleg JL, et al. Outcomes in ambulatory chronic systolic and diastolic heart failure: a propensity score analysis. *Am Heart J*. 2006;152(5):956–966.
17. Stamou SC, White T, Barnett S, et al. Comparisons of cardiac surgery outcomes in Jehovah's versus non-Jehovah's Witnesses. *Am J Cardiol*. 2006;98(9):1223–1225.
18. Toumpoulis IK, Anagnostopoulos CE, Katritsis DG, et al. The impact of preoperative thrombolysis on long-term survival after coronary artery bypass grafting. *Circulation*. 2005;112 (9 suppl):1351–1357.
19. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *J Educ Psychol*. 1974; 66(5):688–701.
20. Imbens GW. Nonparametric estimation of average treatment effects under exogeneity: a review. *Rev Econ Stat*. 2004;86(1): 4–29.
21. Austin PC, Grootendorst P, Anderson GM. A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. *Stat Med*. 2007;26(4):734–753.
22. Austin PC, Grootendorst P, Normand SL, et al. Conditioning on the propensity score can result in biased estimation of common measures of treatment effect: a Monte Carlo study. *Stat Med*. 2007;26(4):754–768.
23. Austin PC. The performance of different propensity score methods for estimating marginal odds ratios. *Stat Med*. 2007; 26(16):3078–3094.
24. Austin PC. The performance of different propensity-score methods for estimating relative risks. *J Clin Epidemiol*. 2008; 61(6):537–545.
25. Austin PC. Some methods of propensity-score matching had superior performance to others: results of an empirical investigation and Monte Carlo simulations. *Biom J*. 2009;51(1): 171–184.
26. Austin PC. Type I error rates, coverage of confidence intervals, and variance estimation in propensity-score matched analyses. *Int J Biostat*. 2009;5(1):Article 13. (doi: 10.2202/1557-4679.1146).
27. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers; 1988.
28. Rosenbaum PR, Rubin DB. The bias due to incomplete matching. *Biometrics*. 1985;41(1):103–116.
29. Rosenbaum P. *Observational Studies*. New York, NY: Springer-Verlag New York; 1995.
30. Austin PC. Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharm Stat*. [published online ahead of print April 10, 2010] (doi: 10.1002/pst.433).
31. Flury BK, Riedwyl H. Standard distance in univariate and multivariate analysis. *Am Stat*. 1986;40(3):249–251.
32. Austin PC. Assessing balance in measured baseline covariates when using many-to-one matching on the propensity-score. *Pharmacoepidemiol Drug Saf*. 2008;17(12): 1218–1225.
33. Casella G, Berger RL. *Statistical Inference*. Belmont, CA: Duxbury Press; 1990.
34. Ury HK. Efficiency of case-control studies with multiple controls per case: continuous or dichotomous data. *Biometrics*. 1975;31(3):643–649.
35. Breslow NE, Day NE. *Statistical Methods in Cancer Research. Vol. 1. The Analysis of Case-Control Studies*. (IARC Scientific Publication no. 32). Lyon, France: International Agency for Research on Cancer; 1980.
36. Ming K, Rosenbaum PR. Substantial gains in bias reduction from matching with a variable number of controls. *Biometrics*. 2000;56(1):118–124.
37. Rosenbaum PR. A characterization of optimal designs for observational studies. *J R Stat Soc Series B*. 1991;53(3): 597–610.
38. Gu XS, Rosenbaum PR. Comparison of multivariate matching methods: structures, distances, and algorithms. *J Comput Graph Stat*. 1993;2(4):405–420.
39. Hansen BB. Full matching in an observational study of coaching for the SAT. *J Am Stat Assoc*. 2004;99(467): 609–618.

Downloaded from https://academic.oup.com/aje/article/172/9/1092/147493 by guest on 18 April 2024